Counterfactual Generative Networks

Year: 2021 | Citations: 140 | Authors: Axel Sauer, Andreas Geiger

---

Abstract

Neural networks are prone to learning shortcuts -- they often model simple correlations, ignoring more complex ones that potentially generalize better. Prior works on image classification show that instead of learning a connection to object shape, deep classifiers tend to exploit spurious correlations with low-level texture or the background for solving the classification task. In this work, we take a step towards more robust and interpretable classifiers that explicitly expose the task's causal structure. Building on current advances in deep generative modeling, we propose to decompose the image generation process into independent causal mechanisms that we train without direct supervision. By exploiting appropriate inductive biases, these mechanisms disentangle object shape, object texture, and background; hence, they allow for generating counterfactual images. We demonstrate the ability of our model to generate such images on MNIST and ImageNet. Further, we show that the counterfactual images can improve out-of-distribution robustness with a marginal drop in performance on the original classification task, despite being synthetic. Lastly, our generative model can be trained efficiently on a single GPU, exploiting common pre-trained models as inductive biases.